DOI:10.13878/j.cnki.jnuist.2017.06.008

黄毅[1,2]　鲍秉坤[1,2]　徐常胜[1,2]

关系挖掘驱动的视频描述自动生成

摘要

视频的自动描述任务是计算机视觉领域的一个热点问题.视频描述语句的生成过程需要自然语言处理的知识,并且能够满足输入(视频帧序列)和输出(文本词序列)的长度可变.为此本文结合了最近机器翻译领域取得的进展,设计了基于编码-解码框架的双层 LSTM 模型.在实验过程中,本文基于构建深度学习框架时重要的表示学习思想,利用卷积神经网络(CNN)提取视频帧的特征向量作为序列转换模型的输入,并比较了不同特征提取方法下对双层 LSTM 视频描述模型的影响.实验结果表明,本文的模型具有学习序列知识并转化为文本表示的能力.

关键词

视频描述;LSTM 模型;表示学习;特征嵌入

中图分类号 TP391.41;TP183

文献标志码 A

0　引言

随着社会网络和在线内容分享服务的迅猛发展,互联网上积累了大量的图像、视频等视觉数据.据统计,每分钟上传 YouTube 视频分享网站的视频长度达到 100 h,而上传至 Flickr 图片分享网站的图片更是多达百万幅.如果能充分理解如此庞大丰富的互联网视觉数据,它们就是一个高价值的信息库,可以进一步为社会服务.然而,为了应对如此大量的视觉信息的收集、分类和处理工作,仅仅依靠人工方法是不够的.这时进行视觉数据的内容理解方面的研究工作就显得尤为重要.

视觉数据的内容理解在计算机视觉和多媒体应用领域已有广泛的研究,包括目标分类、检测和图像描述等.其中视觉数据的描述是近几年新兴的研究方向,主要研究如何自动生成对图片或视频片段的描述性文字,准确表达其所传达的内容.其在人机交互、基于内容的视频搜索、帮助视觉障碍者理解视觉内容等方面都有重要的应用.传统的图像描述方法[1]习惯将其划为 2 个子问题:首先使用图片分类技术,提取图像特征,识别图像中实体、行为和场景;然后再结合从文本语料库挖掘出的统计特征,估计最有可能的主语、动词、宾语和地点的语法结构,最后生成图像的文本描述.用手工设计的语法、根据所识别的内容生成相当有限的描述性句子.这样的方法更多的是关注图像里面有什么,然后总是重复使用描述模型在训练时使用的语句,而对于图像中的物体与物体之间、物体和环境之间的关联及意义并不能给出满意的描述.

显然对图像的自动描述需要更加高级的智能形态.计算机不仅需要识别出图像中的物体,同时必须更加深入理解视觉数据中物体之间以及物体和环境之间的关系,甚至包括一些抽象的属性.图像自动描述研究的突破得益于近年来计算机视觉和自然语言处理领域取得的进步.2012 年,深度卷积神经网络(Deep Convolutional Neural Network,DCNN)在 ImageNet 对象识别挑战赛中首先获得成功[2].紧接着在 2014 年,机器翻译研究获得了巨大的进展,Cho 等[3]研究人员利用一种特殊的循环神经网络——长短期记忆模型(Long Short-Term Memory,LSTM)将源语言的句子编码为一个具有丰富语义知识的向量,然后将这个语意向量作为解码 LSTM 的起始隐藏状态,最后生成目标语言的句子.2015 年,Google Brain 团队的 Vinyals 等[4]

收稿日期 2017-08-28

资助项目 国家自然科学基金(61572503,61432019);北京市自然科学基金(4152053)

作者简介

黄毅,男,硕士生,研究方向为模式识别.huangyi2017@ia.ac.cn

徐常胜(通信作者),男,博士,研究员,博士生导师,主要研究方向是多媒体数据分析及应用.csxu@nlpr.ia.ac.cn

1　中国科学院自动化研究所　模式识别国家重点实验室,北京,100190

2　中国科学院大学,北京,100049

从上述研究中获得了启发,利用 CNN 提取出具有高层语义知识的图片特征,然后将其作为语言生成模型 LSTM 的输入,生成文本序列.在遇见全新场景时,这个模型能够基于图片中物体和环境之间的交互关系,自动生成准确的图像描述,并且使用的自然语言非常流畅.此后,该团队发现对视觉模型和语言生成模型进行端到端的联合训练有利于相互提升效果[5],图片自动描述模型可以生成更精确、更细节化的句子.

而对于开放领域的视频描述,其难点不仅在于难以确定视频中的突出内容,而且很难适当地根据视频前后关系进行事件描述.视频描述模型应允许对可变长度输入序列进行处理,并提供可变长度输出.微软亚洲研究院所提出的方法[6]将二维视觉上的卷积神经网络和三维的动态卷积神经网络结合,并且增加了一种用于探索视觉内容与句子语义之间关系的联合嵌入模型.文献[7-9]都构建了双层 LSTM 的语言生成模型,对视频序列帧编码和文字解码进行联合学习.它们都可以捕捉长期依赖性,能够如同描述静态图片一样很好地描述动态视频.

1 LSTM 视频描述模型

1.1 序列到序列框架

在例如机器翻译、视频描述等许多应用场景中,需要将不同长度的输入序列映射到不同长度的输出序列.用于映射可变长度序列到另一可变长度序列最简单的 RNN 架构最初由 Cho 等[3]提出,之后被使用到机器翻译中,获得了当时最好的结果.研究人员把这种构架称作编码-解码或序列到序列构架.

该构架在给定输入序列 $\boldsymbol{x}^2,\boldsymbol{x}^2,\cdots,\boldsymbol{x}^n$ 的情况下学习生成输出序列 $\boldsymbol{y}^1,\boldsymbol{y}^2,\cdots,\boldsymbol{y}^m$.其工作机制为:作为编码器(Encoder)的 RNN 处理输入序列,输出上下文 C.这个上下文 C 可能是一个概括输入序列 $\boldsymbol{x}^1,\boldsymbol{x}^2,\cdots,\boldsymbol{x}^n$ 的向量或是一个向量序列.之后解码器(Decoder)的 RNN 用 C 作为输入,并产生输出序列.这种构架可以使输入输出序列的长度 n 和 m 彼此不同.2 个 RNN 以最大化 $\log P(\boldsymbol{y}^1,\boldsymbol{y}^2,\cdots,\boldsymbol{y}^m \mid \boldsymbol{x}^1,\boldsymbol{x}^2,\cdots,\boldsymbol{x}^n)$ 为目标共同学习.

1.2 长短期记忆模型

在进行从输入序列到输出序列的映射时,能够很好地利用序列前后之间的关系是循环神经网络(RNN)的一个重要优点.然而在实际操作过程中,标准的 RNN 结构承载长期信息的能力非常有限.给定输入对后续时间步上的隐藏层及输出层的影响,会随着网络的循环而发生指数级的衰减,最后导致网络"忘记"了最早学习到的信息.这种情况在机器学习领域被称作梯度消失(Vanishing Gradient).为了解决这个问题,Hochreiter 等[10]引入自循环的巧妙构思,提出了长短期记忆(Long Short-Term Memory,LSTM)模型.

在传统 RNN 结构中,隐层只有一个向量 $\boldsymbol{h}$ 作为状态变量,这导致 RNN 往往对短期内的输入过于敏感.而 LSTM 网络除了外部的 RNN 循环外,还具有内部的细胞(Cell)自环,用于保存长期状态.如图 1 所示,LSTM 结构包含 3 个门结构,用于控制细胞状态:输入门(Input Gate)控制 LSTM 考虑当前输入 $\boldsymbol{x}_t$ 的程度;忘记门(Forget Gate)控制 LSTM 对先前存储的细胞状态 $\boldsymbol{C}_{t-1}$ 的记忆程度;输出门(Output Gate)用于决定多少记忆转移至隐层 $\boldsymbol{h}_t$.LSTM 的循环公式如下:

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_i\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_i),$$
$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f),$$
$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_o\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_0),$$
$$\widetilde{\boldsymbol{C}}_t=\phi(\boldsymbol{W}_C\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_C),$$
$$\boldsymbol{C}_t=\boldsymbol{f}_t\odot\boldsymbol{C}_{t-1}+\boldsymbol{i}_t\odot\widetilde{\boldsymbol{C}}_t,$$
$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\phi(\boldsymbol{C}_t),$$

上述公式中,σ 表示 Logistic Sigmoid 函数,ϕ 表示双曲正切函数,$\odot$ 运算符表示向量中对应元素相乘.

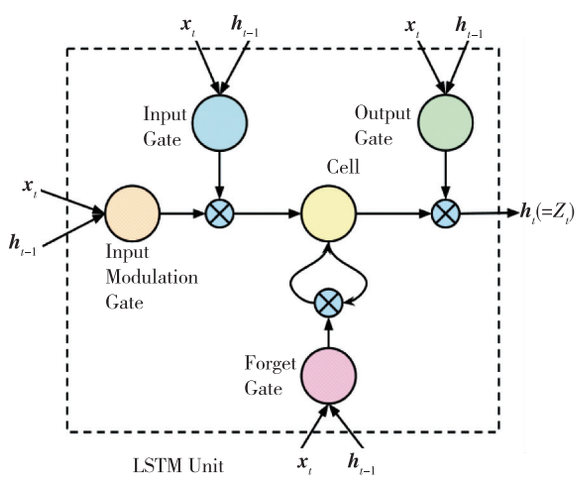


图 1 LSTM 结构

Fig. 1 Architecture of LSTM

LSTM 网络相比简单的循环构架更容易学习长期依赖,其允许网络在较长持续时间内积累信息.一旦中间某些信息被使用,让神经网络选择将其遗忘

的做法确实取得了更好的效果.LSTM 已经在极具挑战性的序列处理任务上已经取得了最先进的水平[11].

1.3 基于双层 LSTM 的序列转换模型

在视频描述任务中,需要处理视频的序列帧,然后输出对应的描述语句.序列到序列的深度学习框架可以很好地满足这个要求.本文使用一个对时间结构比较敏感的双层 LSTM 模型:首先将视频序列帧逐一编码,逐步建立能够有效地编码视频潜在对象、活动和场景的 LSTM 隐层语义表示.一旦读取了视频的所有帧,该模型就会逐词生成一个句子.对于帧的编码和词的解码,都利用平行语料库共同进行学习.这使得该模型具有以下特点:

1) 能够处理不同的输入帧数量;

2) 能够学习和使用视频的时间结构;

3) 能够学习语言模型生成自然语言句子.

LSTM 模型在时间上的展开如图 2 所示.对于 2 个 LSTM,隐层的单元数都设置为 1 000.第 1 层 LSTM 的隐层 $\boldsymbol{h}_t$ 作为第 2 层 LSTM 的输入 $\boldsymbol{x}_t$.在这个模型结构中,第 1 层 LSTM 用于处理视频输入帧序列,第 2 层 LSTM 用于输出单词序列.在前几个时间步中,第 1 层 LSTM 接受序列帧并编码,同时第 2 层 LSTM 接收第 1 层 LSTM 的隐层 $\boldsymbol{h}_t$,然后将其与零向量<pad>连接后进行编码.这段时间,2 个 LSTM 进行编码不进行损失计算.在视频所有帧都被作为输入后,第 2 层 LSTM 接收语句开始标签<BOS>,然后将其循环的隐层解码为单词序列.在解码阶段,通过先前视觉帧序列的编码和已经预测出的单词,对输出的隐层进行下一个单词的预测,最终形成相应的视频描述语句,最后输出结束标志<EOS>.

训练阶段,模型以最大化预测语句的似然值为优化目标.将 θ 作为模型参数,$\boldsymbol{y}=(y_1,y_2,\cdots,y_m)$ 作为输出序列,则优化 θ 的公式可以表示成:

$$\theta^{*}=\underset{\theta}{\operatorname{argmax}}\sum_{t=1}^{m}\log p(y_t\mid h_{n+t-1},y_{y-1};\theta).$$

该似然值的优化采用随机梯度下降法,损失函数的计算只在 LSTM 学习解码阶段进行.在损失值反向传播的过程中,LSTM 学习产生适合输入视频帧序列的隐层表示 $\boldsymbol{h}_t$.第 2 层 LSTM 的输出 $\boldsymbol{z}_t$ 用于预测出现的单词.该模型使用 softmax 函数计算单词 y 在词汇表 V 中的概率分布:

$$p(y\mid z_t)=\frac{\mathrm{e}^{w_y z_t}}{\sum_{y'\in V}\mathrm{e}^{w_{y'}z_t}}.$$

2 基于关系特征嵌入的视频表示

图像的低层视觉特征和高层语义知识存在很大的鸿沟,原始的视频图片帧无法直接作为模型的输入.因此在进行视频到文本的序列转换前,需要得到视频的特征表示.视频帧的表示方式会直接影响 LSTM 描述模型产生的结果,获得视频的有效表示对完成视频描述任务至关重要.

2.1 视频目标特征表示

近年来,深度卷积神经网络(DCNN)在计算机视觉领域的任务中取得了不俗的成绩.DCNN 能够通过将输入图片嵌入到固定长度的向量中,得出原始图像的有效表征.AlexNet、GoogLeNet、VGGNet 和

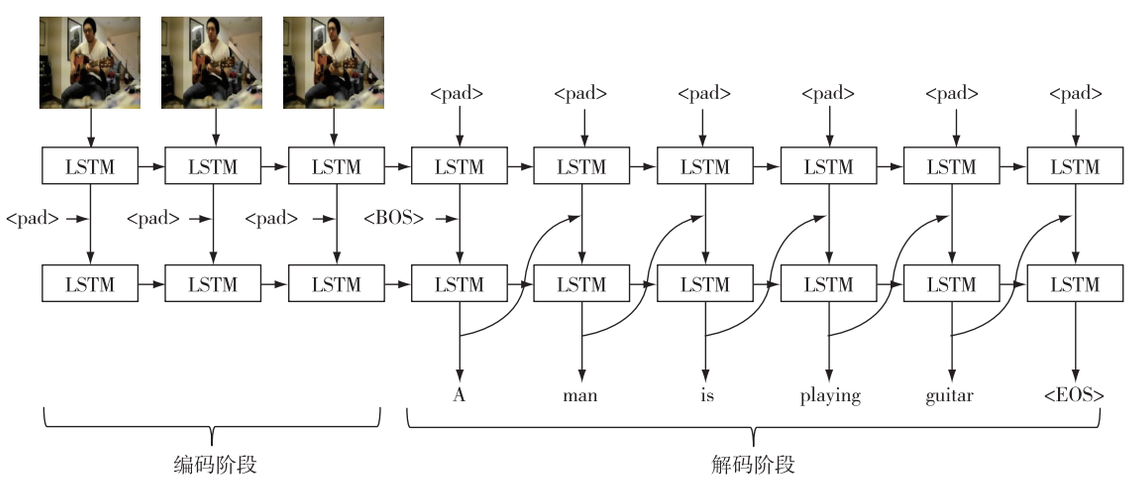


图 2 双层 LSTM 视频描述模型

Fig. 2 Video description model with two-layer LSTM

ResNet 等 CNN 结构都能获得不错的图像目标特征表示. Zhou 等[12] 在 2015 年发现:在 ImageNet 和 Places 基准数据集上训练的深度卷积神经网络中的隐藏单元学习得到的特征通常是可以解释的,它们可以对应人类自然分配的标签.在应用于计算机视觉任务的神经网络中,不是每层的隐藏单元总是能学习出具有简单语言学名称的事物,但有趣的是,这些事物会在深度卷积神经网络的顶层附近出现.因此在视频描述任务中,可以使用 DCNN 预处理视频的序列帧得到视频的序列表示,并将其作为视频描述模型的输入.

2.2 关系特征嵌入

知识表示学习(Knowledge Representation Learning,KRL)源于结构化的文本表示.经典的知识图谱通常以(h,r,t)的三元组形式表示大量结构化的文本信息,其中 h 表示头实体(Head Entity),r 表示实体之间的关系(Relation),t 表示尾实体(Tail Entity).文本 KRL 方法利用基于翻译的技术,将实体和关系投影到连续的低维语义空间中,并将其视为对头部实体和尾部实体之间的关系的翻译操作.受文本 KRL 的启发,视觉 KRL 对包含 2 个对象(对应于文本 KRL 中的"实体")及它们的交互(对应于文本 KRL 中的"关系")的图像区域编码,并映射到语义空间中.

然而,文本 KRL 和视觉 KRL 仅仅专注于从单一模态学习知识表示,而忽略了其他模态的补充信息.虽然现有的知识表示研究取得了很大的成功,但这项研究还需要推广到多模态,使其在实际应用中获得更好的泛化效果.文献[13]提出了一种新型的多模态知识表示学习(Multi-Modal Knowledge Representation Learning,MM-KRL)框架,尝试处理来自文本和视觉 2 个模态的知识,利用文本和视觉 2 个模态进行双增强知识表示学习.该模型通过零点多模态检索来证明它的方法可以将不同的模态的知识投射到共同的知识空间中,而且学习的知识可以表示未知的关系.

在利用视觉模态增强文本知识表示时,定义关系三元组:对于锚定关系 S_i^a,与其相似的视觉模态关系定义为 S_i^p,与其不相似的剩余关系定义为 S_i^n.在训练 LSTM 网络时,除了原有的损失函数,增加了一项包含该三元组的表示:

$$L_{\text{triplet}_i} = \| h(S_i^a) - h(S_i^p) \|_2^2 - \| h(S_i^a) - h(S_i^n) \|_2^2 + \alpha,$$

其中,$h(S_i)$ 是 S_i 的文本关系表示,α 是正相关与负相关对之间的边界值.

在利用文本模态增强视觉知识表示时,使用下式优化 CNN:

$$\min_{W,b} \sum_{i=1}^{M} \sum_{k=1}^{N_i} \| g(v_{ik}) - h(T_i) \|_2^2 + \lambda \| W \|_2^2,$$

其中,M 是训练集中多模态关系的数量,v_{ik} 是关系 i 中的第 k 个图片,N_i 是关系 i 中图片的总数量,g 是图片的视觉知识表示,$h(T_i)$ 是第 i 个文本关系的知识表示.W 和 b 是 CNN 的网络参数,$\lambda \| W \|_2^2$ 是优化的正则化项.整个模型的训练流程如图 3 所示.

与之前的知识表示方法相比,MM-KRL 框架具有以下几个优点:

1)可以自动从网络中有效地挖掘具有结构化的文本和视觉关系的多模态知识;

2)提出的双增强跨模态深度神经网络(Bi-enhanced Cross-modal Deep Neural Network,BC-DNN)

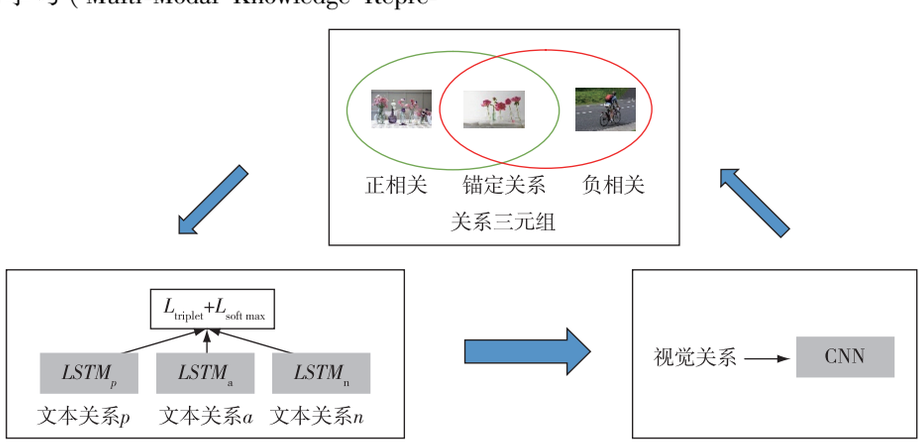


图 3 双增强跨模态深度学习

Fig. 3 Bi-enhanced cross-modal deep learning

能够学习独立于任务和模态的共同知识空间;

3)通过将已经学习得到的知识与已经可见的孤立实体和关系迁移到未知的关系中,能够表示出未知的多模态关系.

3 实验与分析

3.1 视频描述数据集

机器学习中大部分算法都要从数据集(Dataset)上获取经验.近年来神经网络性能的提升,较大的数据集减少了统计泛化对神经网络的挑战的程度.微软视频描述语料库(Microsoft Video Description Corpus,MSVD)项目[14]是工作人员描绘的 YouTube 上单活动短片段(通常不到 10 s)的数据集合.工作人员对 2 000 多个视频给出了多语言描述,其中含有 8 万多个英语描述,并且都包含明确的动作或事件.该数据集之前就在许多行为识别、视频描述任务中被使用.在本实验中,只使用其中的英文描述,并在实验前对文本进行最小化预处理.最终选择 1 200 个视频作为训练集,100 个视频作为验证集.数据集的统计如表 1 所示.

表 1 MSVD 数据集统计

Table 1 Statistics of MSVD dataset

	训练集	验证集
视频数量	1 200	100
参考语句数量	49 120	4 314

3.2 评估指标

对于机器产生的语句,如果人为地评价其好坏,不足之处是其中包含了太多的主观性.近年来,国际上使用一些自动评判标准用于评价机器翻译结果的好坏.目前比较流行、结果较好的有 METEOR 和 CIDER 评价标准.

METEOR 指标[15]最初用于评估机器翻译的结果.该方法基于给定假设语句和一组参考语句对齐程度来计算 METEOR 分数,并且相比较早的评价标准(如 BLEU、ROUGE),METEOR 还包含了同义词匹配.文献[16]表明,当参考语句较少时,METEOR 指标总是比 BLEU、ROUGE 更好,与人为评价结果有较高的相关性.METEOR 指标后来也被用于图片描述领域的结果评价,并且也有很好的表现[17].

CIDER 评价标准[16]不同于机器翻译的评价标准,该标准是在计算机视觉与模式识别大会上首次提出的.CIDER 是针对图像描述问题的评价标准.之前评价标准关注的是结果与参考语句的相关性,而 CIDER 更加关注计算机自动生成的句子与人为书写的句子的相似性,而且在含有大量参考语句的评价过程中,CIDER 在匹配度上的表现要好于 METEOR 标准.

3.3 实验过程

本文使用 CAFFE(Convolutional Architecture for Fast Feature Embedding)框架搭建深度神经网络模型[18],并对神经网络进行训练和测试.

实验中,作为输出目标的单词序列采用独热编码(One-Hot Encoding),第 2 层的 LSTM 的输出 $\boldsymbol{h}_t^2$ 的维数设置为词汇表中单词的个数 N,然后对 $\boldsymbol{h}_t^2$ 应用 softmax 函数,归一化概率并最终确定输出的词汇.在 LSTM 自循环过程中,将输出单词向量 $\boldsymbol{h}_t^2$ 嵌入到较低的 500 维空间中,并与第 1 层的 LSTM 的输出 $\boldsymbol{h}_t^1$ 连接,共同构成第 2 层 LSTM 的输入.

针对视频帧序列的知识表示,本文进行了 2 个实验:

1) 使用 ResNet-152[19]提取视频帧的知识表示.实验过程中,移除了网络最后一个全连接层,将最后一个池化层的 2 048 维向量作为视频帧的知识表示,之后将其嵌入到一个 500 维空间中,作为视频描述模型中第 1 层 LSTM 的输入 $\boldsymbol{x}_t^1$.

2) 利用 BC-DNN 提取图像的关系特征,将其最后一个全连接层 fc9 的 128 维特征向量与 ResNet-152 中得到的 2 048 维的特征向量连接,并嵌入到一个 500 维空间中,作为视频描述模型中第 1 层 LSTM 的输入 $\boldsymbol{x}_t^1$.

其中,从高维映射到低维的连接权重与双层 LSTM 网络共同训练.在整个训练过程中,综合考虑训练的显存消耗和模型处理视频的长度的能力,我们将双层 LSTM 视频描述模型展开成 80 个时间步长,训练批量设置为 32.在视频帧序列输入结束后,使用 0 填充后续的输入.在切分视频时,每隔 10 帧提取一个样本,对于切分后的 MSVD 数据集中的视频,都可以在 80 个时间步长内编码完成并解码输出词序列.

3.4 结果分析

如表 2 所示的是视频描述模型在 MSVD 数据集上的实验结果.其中第 1 行是文献[9]中利用 VGGNet 提取视频特征训练 LSTM 描述模型的实验结果.本文使用基于 ResNet-152 的特征提取和 ResNet-152+BC-DNN 的连接特征,在 MSVD 数据集上进行视频描述实验.

表 2　不同特征表示下视频描述的评价结果

Table 2　Evaluation results of the video description by different feature representations

模型	评估指标	
	METEOR	CIDER
VGGNet[9]	29.0	52.6
ResNet-152	30.9	50.8
ResNet+BC-DNN	29.9	56.2

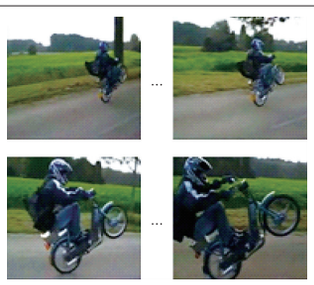
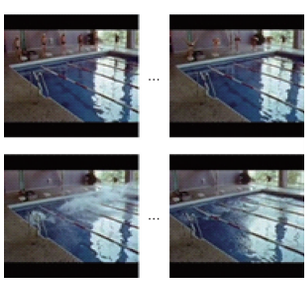
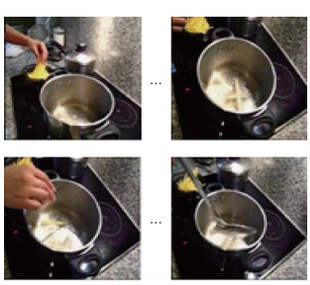
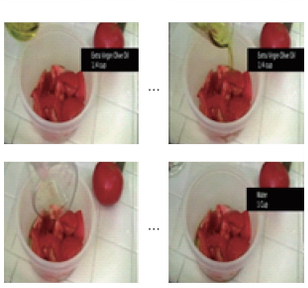
基于 ResNet-152 的视频特征提取在 METEOR 指标上从 VGGNet 的 29.0 提高到了 30.9,CIDER 指标上比 VGGNet 略有下降.因为 ResNet-152 相比 VGGNet 对图像中对象特征有更丰富的表示,在关注描述相关性的 METEOR 指标下,模型效果明显提高.但是对象关系特征没有提高,使关注描述结果与人为描述相似性的 CIDER 成绩下降.

在 ResNet-152 特征中嵌入 BC-DNN 提取的关系特征后,视频描述模型在 METEOR 和 CIDER 上的表现都比 VGGNet 情况有了提高,尤其是在关注与人为描述相似性下的 CIDER 得分从 50.8 提高到了 56.2.但是在描述相关性评价的 METEOR 上比单纯使用 ResNet-152 时的得分略有下降.

表 3 所示是 3 种不同特征表示方法下描述结果的例子.相比另外 2 种表示,ResNet+BC-DNN 情况下的描述结果突出了视频中对象之间的关系.例如表 3

表 3　不同特征表示下的视频描述结果示例

Table 3　Some results of the video description by different feature representations

视频片段	自动描述结果	人为真实描述
	**LSTM(VGGNet)**:A man is riding a motorcycle. **LSTM(ResNet-152)**:A man is riding a bicycle. **LSTM(ResNet-152+BC-DNN)**:A man is doing stunts on a motorcycle.	①The man is doing moped tricks. ②A person is doing a motorcycle trick. ③A person is doing a wheelie on a bicycle.
	**LSTM(VGGNet)**:A man is jumping in a beach. **LSTM(ResNet-152)**:A man is jumping. **LSTM(ResNet-152+BC-DNN)**:A man is jumping into the water.	①People are jumping in the water. ②People dive into a pool. ③People getting into a swimming pool.
	**LSTM(VGGNet)**:A man is cooking. **LSTM(ResNet-152)**:A man is cooking the rice-aroni. **LSTM(ResNet-152+BC-DNN)**:A person puts a piece of cheese into a skillet.	①A person is frying a food. ②Someone is frying potatoes. ③A cook puts noodles into some boiling water.
	**LSTM(VGGNet)**:A woman is pouring some meat. **LSTM(ResNet-152)**:A man is adding oil to a pot. **LSTM(ResNet-152+BC-DNN)**:A man pours some sauce from a container into a glass bottle.	①A man is mixing something into tomatoes. ②The man is pouring oil on the tomatoes. ③A man is adding oil to a bowl of tomatoes.

中第 1 行,ResNet+BC-DNN 特征识别出了人与车之间不是简单的骑行关系,而是人在车上做特技.

4 结束语

为视频中的事件生成自然语言描述具有多种实际应用.近年来,研究者们对静态图像和视频描述的兴趣日益激增.为了使用自然语言自动描述更广泛的普通视频,需要实现语言和视觉语意更深层次的整合.视频自动描述技术应该具备识别值得描述的突出事件的能力,并且应该能够适当地描述具有大量不同动作、对象、场景和其他属性的各种视频内容.基于深度神经网络的视频描述模型在这个方向上迈出了重要的一步.

本文使用序列到序列建模的方法构建视频模型描述,模型首先在编码阶段按顺序读取视频帧,然后解码按序生成文字.该模型允许处理可变长度的输入和输出,同时可以对时间结构建模.它将视频视为一种“语言”,并采用机器翻译的方法将视频翻译成文本,能够直接从视频和句子对中学习值得描述的显著对象.文本的模型在 MSVD 数据集上的实验获得了较好的表现.基于 ResNet-152 的视频特征使视频表示出更丰富的内容,描述结果的相关性更好.基于 ResNet-152+BC-DNN 的特征提取在视频表示中加入了关系特征,很大地提高了描述结果与人为描述的相似性.

参考文献

References

[1] Thomason J,Venugopalan S,Guadarrama S,et al.Integrating language and vision to generate natural language descriptions of videos in the wild[C]//International Conference on Computational Linguistics,2014:1218-1227

[2] Krizhevsky A,Sutskever I,Hinton G E.ImageNet classification with deep convolutional neural networks[C]//International Conference on Neural Information Processing Systems,2012:1097-1105

[3] Cho K, Merrienboer B V, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv e-print, 2014, arXiv:1406. 1078

[4] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[J]. arXiv e-print, 2014, arXiv:1411. 4555

[5] Vinyals O, Toshev A, Bengio S, et al. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 652-663

[6] Pan Y W, Mei T, Yao T, et al. Jointly modeling embedding and translation to bridge video and language[C]//IEEE Conference on Computer Vision and Pattern Recognition,2016:4594-4602

[7] Donahue J,Hendricks L A,Rohrbach M,et al.Long-term recurrent convolutional networks for visual recognition and description[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(4): 677-691

[8] Venugopalan S,Xu H,Donahue J,et al.Translating videos to natural language using deep recurrent neural networks[J].arXiv e-print,2015,arXiv:1412. 4729

[9] Venugopalan S,Rohrbach M,Donahue J,et al.Sequence to sequence:Video to text[C]//IEEE International Conference on Computer Vision,2015:4534-4542

[10] Hochreiter S, Schmidhuber J. Long short-term memory[J].Neural Computation,1997,9(8):1735-1780

[11] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[J]. arXiv e-print, 2014, arXiv:1409. 3215

[12] Zhou B L,Khosla A,Lapedriza A,et al.Object detectors emerge in deep scene CNNs[J].arXiv e-print,2015, arXiv:1412. 6856

[13] Nian F D, Bao B K, Li T, et al. Multi-modal knowledge representation learning via webly-supervised relationships mining[C]//ACM International Conference on Multimedia,2017(accepted)

[14] Chen D L,Dolan W B.Collecting highly parallel data for paraphrase evaluation[C]//Meeting of the Association for Computational Linguistics:Human Language Technologies,2011:190-200

[15] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language[C]//Workshop on Statistical Machine Translation, 2014: 376-380

[16] Vedantam R,Zitnick C L,Parikh D.CIDEr:Consensus-based image description evaluation[J].IEEE Conference on Computer Vision and Pattern Recognition, 2015: 4566-4575

[17] Elliott D,Keller F.Comparing automatic evaluation measures for image description[C]//Meeting of the Association for Computational Linguistics,2013:452-457

[18] Jia Y Q,Shelhamer,Donahue J,et al.Caffe:Convolutional architecture for fast feature embedding[J].arXiv e-print, 2014,arXiv:1408. 5093

[19] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[J]. arXiv e-print, 2015, arXiv:1512. 03385

南京信息工程大学学报(自然科学版),2017,9(6):642-649
Journal of Nanjing University of Information Science and Technology(Natural Science Edition),2017,9(6):642-649

Video description based on relationship feature embedding

HUANG Yi[1,2] BAO Bingkun[1,2] XU Changsheng[1,2]
1 National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190
2 University of Chinese Academy of Sciences, Beijing 100049

Abstract Video description has received increased interest in the field of computer vision. The process of generating video descriptions needs the technology of natural language processing, and the capacity to allow both the lengths of input (sequence of video frames) and output (sequence of description words) to be variable. To this end, this paper uses the recent advances in machine translation, and designs a two-layer LSTM (Long Short-Term Memory) model based on the encoder-decoder architecture. Since the deep neural network can learn appropriate representation of input data, we extract the feature vectors of the video frames by convolution neural network (CNN) and take them as the input sequence of the LSTM model. Finally, we compare the influences of different feature extraction methods on the LSTM video description model. The results show that the model in this paper is able to learn to transform sequence of knowledge representation to natural language.

Key words video description; LSTM model; representation learning; feature embedding